# Epistemic foundations for backward induction: an overview

Andrés Perea

# Epistemic Foundations for Backward Induction: An Overview

RM/06/036

JEL code: C72

**M**aastricht research school of **E**conomics
of **TE**chnology and **OR**ganizations

Universiteit Maastricht
Faculty of Economics and Business Administration
P.O. Box 616
NL - 6200 MD Maastricht

phone : ++31 43 388 3830
fax : ++31 43 388 4873

# Epistemic Foundations for Backward Induction: An Overview

Andrés Perea*
Maastricht University

April 2006

**Abstract**

In this survey we analyze, and compare, various sufficient epistemic conditions for backward induction that have been proposed in the literature. To this purpose we present a simple epistemic base model for games with perfect information, and translate the different models into the language of this base model. As such, we formulate the various sufficient conditions for backward induction in a uniform language, which enables us to explictly analyze their differences and similarities.

*Keywords*: Backward induction, epistemic game theory, belief revision.

## 1. Introduction

Backward induction constitutes one of the oldest concepts in game theory. Its algorithmic definition, which goes back at least to Zermelo (1913), seems so natural at first sight that one might be tempted to argue that every player "should" reason in accordance with backward induction in every game with perfect information. However, on a decision theoretic level the concept is no longer as uncontroversial as it may seem. The problem is that the backward induction algorithm, when applied from a certain decision node on, completely ignores the history that has led to this decision node, as it works from the terminal nodes towards this decision node. At the same time, the beliefs that the player at this decision node has about his opponents' future behavior may well be affected by the history he has observed so far. For instance, a player who observes that an opponent has not chosen in accordance with backward induction in the past may have a valid reason to believe that this same opponent will continue this pattern in the game that lies ahead. However, such belief revision policies are likely to

*Address: Department of Quantitative Economics, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. E-mail: a.perea@ke.unimaas.nl. Web: http://www.personeel.unimaas.nl/a.perea/. Tel: +31-43-3883922.

induce choices that contradict backward induction. We therefore need to impose some non-trivial conditions on the players' belief revision policies in order to arrive at backward induction.

During the last decade or so, the game-theoretic literature has provided us with various epistemic models for dynamic games in which sufficient epistemic conditions for backward induction have been formulated. The objective of this survey is to discuss these conditions individually, and to explicitly compare the different conditions with each other. The latter task is particularly difficult since the literature exhibits a large variety of epistemic models, each with its own language, assumptions and epistemic operators. Some models are syntactic while others are semantic, and among the semantic models some are based on the notion of states of the world while others use types instead. As to the epistemic operators, some models apply knowledge operators while others use belief operators, and there is also a difference with respect to the "timing" of these operators. Are players entitled to revise their knowledge or belief during the course of the game, and if so, at which instances can they do so? Different models provide different answers to these, and other, questions.

As to overcome these problems we present in Section 2 an epistemic base model, which will be used as a kind of "uniform language" into which all other models can be translated. In Section 3 we then provide for each of the papers to be discussed a brief description of the model, followed by a translation of its epistemic conditions for backward induction in terms of our base model. By doing so we formulate all sufficient conditions for backward induction in the same language, which makes it possible to explictly analyze the differences and similarities between the various conditions.

Finally, a word about the limitations of this paper. In this survey, we restrict attention to epistemic conditions that lead to the *backward induction strategies for all players.* There are alternative models that lead to the *backward induction outcome,* but not necessarily to the backward induction strategy for each player. For instance, Battigalli and Siniscalchi (2002) and Brandenburger, Friedenberg and Keisler (2004) provide epistemic models for extensive form rationalizability (Pearce (1984), Battigalli (1997)) and iterated maximal elimination of weakly dominated strategies, respectively, which always lead to the backward induction outcome in every generic game with perfect information, but not necessarily to the backward induction strategy profile. We also focus exclusively on sufficient conditions that apply to *all* generic games with perfect information. There are other interesting papers that deal with the logic of backward induction in *specific* classes of games, such as Rosenthals's centipede game (Rosenthal (1981)) and the finitely repeated prisoner's dilemma. See, among others, Binmore (1987), Stalnaker (1996), Aumann (1998), Rabinowicz (1998), Broome and Rabinowicz (1999), Carroll (2000) and Priest (2000). We shall, however, not discuss these papers here. Even with the limitations outlined above, we do not claim to offer an exhaustive list of epistemic models for backward induction. We do believe, however, that the list of models treated here will give the reader a good impression of the various epistemic conditions for backward induction that exist in the literature.

## 2. An Epistemic Base Model

### 2.1. Games with Perfect Information

A dynamic game is said to be with *perfect information* if every player, at each instance of the game, observes the opponents' moves that have been made until then. Formally, an *extensive form structure* $\mathcal{S}$ *with perfect information* consists of a finite game tree, a finite set $I$ of players, for every player $i$ a finite set $H_i$ of information sets, for every information set $h_i \in H_i$ a finite set $A(h_i)$ of available actions, and a finite set $Z$ of terminal nodes. By $h_0$ we denote the beginning of the game, and we write $H_i^* = H_i \cup \{h_0\}$. Perfect information is modeled by the assumption that each decision node by itself constitutes an information set. By $A$ we denote the set of all actions, whereas $H$ denotes the collection of all information sets. We assume that no chance moves occur. The definition of a strategy we shall employ coincides with the concept of a *plan of action,* as discussed in Rubinstein (1991). The difference with the usual definition is that we require a strategy only to prescribe an action at those information sets that the same strategy does not avoid. Formally, let $\tilde{H}_i \subseteq H_i$ be a collection of player $i$ information sets, not necessarily containing all information sets, and let $s_i : \tilde{H}_i \to A$ be a mapping prescribing at every $h_i \in \tilde{H}_i$ some available action $s_i(h_i) \in A(h_i)$. For a given information set $h \in H$, not necessarily belonging to player $i$, we say that $s_i$ *avoids* $h$ if there is some $h_i \in \tilde{H}_i$ on the path to $h$ at which the prescibed action $s_i(h_i)$ deviates from the path to $h$. Such a mapping $s_i : \tilde{H}_i \to A$ is called a *strategy* for player $i$ if $\tilde{H}_i$ is exactly the collection of player $i$ information sets not avoided by $s_i$. Obviously, every strategy $s_i$ can be obtained by first prescribing an action at all player $i$ information sets, that is, constructing a strategy in the classical sense, and then deleting from its domain those player $i$ information sets that are avoided by it. For a given strategy $s_i \in S_i$, we denote by $H_i(s_i)$ the collection of player $i$ information sets that are not avoided by $s_i$. Let $S_i$ be the set of player $i$ strategies. For a given information set $h \in H$ and player $i$, we denote by $S_i(h)$ the set of player $i$ strategies that do not avoid $h$. Then, it is clear that a profile $(s_i)_{i \in I}$ of strategies reaches an information set $h$ if and only if $s_i \in S_i(h)$ for all players $i$.

### 2.2. Preferences, Beliefs and Types

The basic assumption is that every player has a strict preference relation over the terminal nodes, and holds at each of his information sets a conditional belief about the opponents' strategy choices and preference relations. In particular, we allow for the fact that players may revise their beliefs about the opponents' preferences as the game proceeds. In order to keep our model as "weak" as possible, we assume that this conditional belief can be expressed by a *set* of opponents' strategies and preference relations. This set represents the strategies and preference relations that the player deems *possible* at his information set. We thus do not consider probabilities, and it is therefore sufficient to specify the players' *ordinal* preferences over terminal nodes. Not only does a player hold first-order conditional beliefs about the opponents' choices and preferences, he also holds second-order conditional beliefs about the opponents' possible first-order beliefs

at each of his information sets. A second-order belief may thus contain expressions of the form "player $i$ considers it possible at information set $h_i$ that player $j$ considers it possible at information set $h_j$ that player $k$ chooses strategy $s_k$ and has preference relation $P_k$". Recursively, one may define higher-order conditional beliefs for the players. A possible way to represent such hierarchies of conditional beliefs is by means of the following model.

**Definition 2.1.** *(Epistemic base model) Let $\mathcal{S}$ be an extensive form structure with perfect information. An epistemic base model for $\mathcal{S}$ is a tuple*

$$\mathcal{M} = ((T_i)_{i \in I}, (B_{ij})_{j \neq i}, (P_i)_{i \in I})$$

*where*
*(1) $T_i$ is a set of types for player $i$;*
*(2) $P_i$ is a function that assigns to every $t_i \in T_i$ some strict preference relation $P_i(t_i)$ over the terminal nodes;*
*(3) $B_{ij}$ is a function that assigns to every $t_i \in T_i$ and every information set $h_i \in H_i^*$ some subset $B_{ij}(t_i, h_i) \subseteq S_j(h_i) \times T_j$.*

From an epistemic base model, conditional beliefs of any order can be *derived*. For instance, type $t_i$'s belief at $h_i$ about player $j$'s choice is given by the projection of $B_{ij}(t_i, h_i)$ on $S_j$. Let $B_{ij}(t_i, h_i|S_j)$ denote this projection, and let $B_{ij}(t_i, h_i|T_j)$ denote its projection on $T_j$. Then, type $t_i$'s belief at $h_i$ about player $j$'s belief at $h_j$ about player $k$'s choice is given by

$$\bigcup_{t_j \in B_{ij}(t_i, h_i|T_j)} B_{jk}(t_j, h_j|S_k).$$

In a similar fashion, higher order beliefs can be derived.

### 2.3. Common Belief

Let $\mathcal{M} = ((T_i)_{i \in I}, (P_i)_{i \in I}, (B_{ij})_{i \neq j})$ be an epistemic base model, and $E \subseteq \cup_{i \in I} T_i$ a set of types, or *event.* We say that type $t_i$ *believes in* $E$ at information set $h_i \in H_i^*$ if $B_{ij}(t_i, h_i|T_j) \subseteq E$ for all $j \neq i$. We say that $t_i$ *initially believes in* $E$ if $t_i$ believes in $E$ at $h_0$. Common belief in the event $E$ is defined by the following recursive procedure:

$$B_i^1(E) = \{t_i \in T_i \mid t_i \in E\}$$

for all $i \in I$, and

$$B_i^{k+1}(E) = \{t_i \in B_i^k(E) \mid B_{ij}(t_i, h_i|T_j) \subseteq B_j^k(E) \text{ for all } j \neq i \text{ and all } h_i \in H_i^*\}$$

for all $i \in I$ and all $k \geq 2$.

**Definition 2.2.** *(Common belief) A type $t_i \in T_i$ is said to respect common belief in the event $E$ if $t_i \in B_i^k(E)$ for all $k$.*

Hence, $t_i$ respects common belief in $E$ if $t_i$ belongs to $E$, believes throughout the game that opponents' types belong to $E$, believes throughout the game that opponents believe throughout the game that the other players' types belong to $E$, and so on. Common initial belief in the event $E$ is defined as follows:

$$IB_i^1(E) = \{t_i \in T_i \mid t_i \in E\}$$

for all $i \in I$, and

$$IB_i^{k+1}(E) = \{t_i \in IB_i^k(E) \mid B_i(t_i, h_0|T_j) \subseteq IB_j^k(E) \text{ for all } j \neq i\}$$

for all $i \in I$ and all $k \geq 2$.

**Definition 2.3.** *(Common initial belief) A type $t_i \in T_i$ is said to respect common initial belief in the event $E$ if $t_i \in IB_i^k(E)$ for all $k$.*

### 2.4. Belief in the Opponents' Rationality

All the epistemic foundations for backward induction to be discussed here make assumptions about the beliefs that players have about the rationality of their opponents. More precisely, all foundations require that players *initially* believe that each opponent chooses rationally at every information set. However, the various foundations differ as to how players would *revise* their beliefs upon observing that their initial belief about the opponents was incorrect. In order to translate these different belief revision procedures into our base model, we need the following definitions.

We first define what it means that a strategy is rational for a type at a given information set. For a strategy $s_i$, an opponents' strategy profile $s_{-i} \in \times_{j \neq i} S_j$, and an information set $h_i \in H_i$, let $z(s_i, s_{-i}|h_i)$ be the terminal node that would be reached from $h_i$ if $(s_i, s_{-i})$ were to be executed by the players.

**Definition 2.4.** *(Rationality at an information set) A strategy $s_i$ is rational for type $t_i$ at information set $h_i \in H_i(s_i)$ if there is no $s_i' \in S_i(h_i)$ such that $P_i(t_i)$ ranks $z(s_i', s_{-i}|h_i)$ strictly over $z(s_i, s_{-i}|h_i)$ for all $s_{-i} \in \times_{j \neq i} B_{ij}(t_i, h_i|S_j)$.*

We shall now define various restrictions on the beliefs that players have about the opponents' rationality. We need one more definition to this purpose. For a given type $t_i \in T_i$, information set $h_i \in H_i^*$, and some opponent's information set $h_j \in H_j$ following $h_i$, we say $t_i$ *believes $h_j$ to be reached from $h_i$* if $B_{ik}(t_i, h_i|S_k) \subseteq S_k(h_j)$ for all $k \neq i$.

**Definition 2.5.** *(Belief in the opponents' rationality)*
*(1) Type $t_i$ believes at information set $h_i \in H_i^*$ that player $j$ chooses rationally at information set $h_j \in H_j$ if for every $(s_j, t_j) \in B_{ij}(t_i, h_i)$ it is true that $s_j$ is rational for $t_j$ at $h_j$.*
*(2) Type $t_i$ initially believes in rationality at all information sets if $t_i$ believes at $h_0$ that every opponent $j$ chooses rationally at all $h_j \in H_j$.*
*(3) Type $t_i$ always believes in rationality at all future information sets if $t_i$ believes at every $h_i \in H_i^*$ that every opponent $j$ chooses rationally at every $h_j \in H_j$ that follows $h_i$.*
*(4) Type $t_i$ always believes in rationality at future information sets that are believed to be reached if $t_i$ believes at every $h_i \in H_i^*$ that every opponent $j$ chooses rationally at all those $h_j \in H_j$ following $h_i$ which $t_i$ believes to be reached from $h_i$.*
*(5) Type $t_i$ always believes in rationality at all future and parallel information sets if $t_i$ believes at every $h_i \in H_i^*$ that every opponent $j$ chooses rationally at every $h_j \in H_j$ that does not precede $h_i$.*
*(6) Type $t_i$ always believes in rationality at all information sets if $t_i$ believes at every $h_i \in H_i^*$ that every opponent $j$ chooses rationally at every $h_j \in H_j$.*

Combined with the definition of common belief, we may thus construct phrases as "type $t_i$ respects common belief in the event that all types initially believe in rationality at all information sets". Some of the epistemic foundations for backward induction, however, use a condition that cannot be expressed in this form, since it relies on a notion that is different from common belief. In order to formalize this condition, we consider the following recursive procedure:

$$FBSR_i^1(h_i) = \{t_i \in T_i \mid t_i \text{ believes at } h_i \text{ that every } j \neq i \text{ chooses rationally at all } h_j \text{ that follow } h_i\}$$

for all $i \in I$ and all $h_i \in H_i^*$, and

$$FBSR_i^k(h_i) = \{t_i \in T_i \mid B_{ij}(t_i, h_i | T_j) \subseteq FBSR_j^{k-1}(h_j) \text{ for all } j \neq i \text{ and all } h_j \text{ that follow } h_i\}$$

for all $i \in I$, $h_i \in H_i^*$ and $k \geq 2$.

**Definition 2.6.** *(Forward belief in substantive rationality) A type $t_i$ is said to respect forward belief in substantive rationality if $t_i \in FBSR_i^k(h_i)$ for all $k$ and all $h_i \in H_i^*$.*

That is, $t_i$ respects forward belief in substantive rationality if $t_i$ (1) always believes that every opponent is rational at every future information set, (2) always believes that every opponent, at every future information set, believes that every opponent is rational at every future information set, (3) always believes that every opponent, at every future information set, believes that every

opponent, at every future information set, believes that every opponent is rational at every future information set, and so on.

We also present a weaker version of forward belief in rationality, which we call forward belief in *material* rationality. Let $H_j(t_i, h_i)$ be the set of those player $j$ information sets $h_j$ following $h_i$ which $t_i$ believes to be reached from $h_i$. Consider the following recursive procedure:

$$FBMR_i^1(h_i) = \{t_i \in T_i \mid t_i \text{ believes at } h_i \text{ that every } j \neq i \text{ chooses rationally at all } h_j \text{ in } H_j(t_i, h_i)\}$$

for all $i \in I$ and all $h_i \in H_i^*$, and

$$FBMR_i^k(h_i) = \{t_i \in T_i \mid B_{ij}(t_i, h_i|T_j) \subseteq FBMR_j^{k-1}(h_j) \text{ for all } j \neq i \text{ and all } h_j \text{ in } H_j(t_i, h_i)\}$$

for all $i \in I$, $h_i \in H_i^*$ and $k \geq 2$.

**Definition 2.7.** *(Forward belief in material rationality) A type $t_i$ is said to respect forward belief in material rationality if $t_i \in FBMR_i^k(h_i)$ for all $k$ and all $h_i \in H_i^*$.*

The crucial difference with forward belief in substantive rationality is thus that a type only believes his opponents to choose rationally at future information sets *which he believes to be reached.* And a type only believes the opponents' types to believe so at future information sets which he believes to be reached, and so on.

## 3. Epistemic Foundations for Backward Induction

In this section we provide an overview of various epistemic foundations that have been offered in the literature for backward induction. A comparison between these foundations is difficult, since the models used by these foundations differ on many aspects.

A first important difference lies in the way the players' beliefs about the opponents are expressed. Some models express the players' beliefs *directly* by means of logical propositions in some formal language. Other models represent the players' beliefs *indirectly* by a set of states of the world, and assign to each state and every player some strategy choice for this player, together with a belief that the player holds at this state about the state of the world. From this model we can derive the higher-order beliefs that players hold about the opponents' choices and beliefs. There are yet some other models that represent the players' beliefs indirectly by means of types, and assign to every type some belief about the other players' choices and types. Similarly to the previous approach, the players' higher-order beliefs can be derived from this model. We refer to these three approaches as the *syntactic model,* the *state-based semantic model* and the *type-based syntactic model.* Note that our base model from the previous section belongs to the

last category. This choice is somewhat arbitrary, since we could as well have chosen a syntactic or state-based semantic base model.

Even within the state-based semantic model, the various papers differ on the precise formalization of the beliefs that players have about the state of the world. Similarly, within the type-based model different papers use different belief operators expressing the players' beliefs about the opponents' choices and types.

Finally, some models impose additional conditions on the extensive form structure, such as one information set per player, or the presence of only two players, whereas other papers do not.

In spite of these differences, all foundations have two aspects in common. First, all models provide a theorem, say Theorem A, which gives a sufficient condition for backward induction. Hence, Theorem A states that if player $i$'s belief revision procedure about the other players' choices, preferences and beliefs satisfies some condition **BR**, then his unique optimal choice is his backward induction choice. Secondly, all models guarantee that this sufficient condition **BR** is possible. That is, each paper provides a second result, say Theorem B, which states that for every player $i$ there is some model in which player $i$'s belief revision procedure satisfies condition **BR**. As we will see, the various foundations differ in the sufficient condition **BR** that is being employed.

In order to explicitly compare the different foundations for backward induction, we attempt to "translate" the various conditions **BR** used by the different models in a unified language, namely the language of our base model. By doing so, we translate the Theorems A and B used by the various foundations into the following standardized form:

**Theorem A:** Let $\mathcal{S}$ be an extensive form structure with perfect information, and let $\mathcal{M} = ((T_i)_{i \in I}, (P_i)_{i \in I}, (B_{ij})_{i \neq j})$ be an epistemic base model for $\mathcal{S}$. Let $(\tilde{P}_i)_{i \in I}$ be a profile of strict preference relations over the terminal nodes. If type $t_i \in T_i$ has preference relation $\tilde{P}_i$, and if $t_i$'s conditional belief vector about the opponents' strategy choices and types satisfies condition **BR,** then there is a unique strategy that is rational for $t_i$ at all information sets, namely his backward induction strategy in the game given by $(\tilde{P}_i)_{i \in I}$.

**Theorem B:** Let $\mathcal{S}$ be an extensive form structure with perfect information, and let $i$ be a player. Then, there is some epistemic base model $\mathcal{M} = ((T_i)_{i \in I}, (P_i)_{i \in I}, (B_{ij})_{i \neq j})$ for $\mathcal{S}$ and some type $t_i \in T_i$ such that $t_i$'s conditional belief vector satisfies **BR.**

In the overview that follows, we provide a brief description of every model, identify the condition **BR** that is being used, and explain how this condition may be translated into the language of our base model. The models are put in alphabetical order.

### 3.1. Asheim's Model

Asheim (2002) uses a type-based semantic model, restricted to the case of two players, in which the players' beliefs are modelled by lexicographic probability distributions. Formally, an Asheim-

model is given by a tuple

$$\mathcal{M} = ((T_i)_{i \in I}, (\lambda_i)_{i \in I}, (v_i)_{i \in I})$$

where $T_i$ is a finite set of types, $v_i$ is a function that assigns to every $t_i$ some vNM-utility function $v_i(t_i)$ over the set of terminal nodes, and $\lambda_i$ is a function that assigns to every type $t_i$ some lexicographic probability system $\lambda_i(t_i)$ on $S_j \times T_j$ with full support on $S_j$. Such a *lexicographic probability system* (or LPS) $\lambda_i(t_i)$ is given by a vector $(\lambda_i^1(t_i), ..., \lambda_i^{K_i(t_i)}(t_i))$ of probability distributions on $S_j \times T_j$. The interpretation is that $\lambda_i^1(t_i), ..., \lambda_i^{K_i(t_i)}(t_i)$ represent different degrees of beliefs, and that the $k$-th degree belief $\lambda_i^k(t_i)$ is infinitely more important than the $(k+1)$-th degree belief $\lambda_i^{k+1}(t_i)$, without completely discarding the latter. The LPS $\lambda_i(t_i)$ induces in a natural way first-order conditional beliefs about player $j$'s choices, as defined in our base model. Namely, for every $h_i \in H_i^*$, let $k_i(t_i, h_i)$ be the first $k$ such that $\lambda_i^k(t_i)$ assigns positive probability some strategy $s_j \in S_j(h_i)$, and let $\hat{B}_{ij}(t_i, h_i) \subseteq S_j(h_i)$ be the set of strategies in $S_j(h_i)$ to which $\lambda_i^{k_i(t_i,h_i)}(t_i)$ assigns positive probability. Then, $t_i$ induces the conditional belief vector $(\hat{B}_{ij}(t_i, h_i))_{h_i \in H_i^*}$ about player $j$'s strategy choice. For every $h_i$, let $\hat{T}_{ij}(t_i, h_i) \subseteq T_j$ be the set of types to which $\lambda_i^{k_i(t_i,h_i)}(t_i)$ assigns positive probability. Then, the induced second-order belief of $t_i$ at $h_i$ about player $j$'s belief at $h_j$ about player $i$'s choice is given by the union of the sets $\hat{B}_{ji}(t_j, h_j)$ with $t_j \in \hat{T}_{ij}(t_i, h_i)$. Similarly, higher-order beliefs about strategy choices can be derived from Asheim's model.

In Asheim's model, a strategy $s_i$ is called rational for type $t_i \in T_i$ at information set $h_i$ if $s_i$ is optimal with respect to the utility function $v_i(t_i)$ and the LPS $\lambda_i(t_i|h_i)$, where $\lambda_i(t_i|h_i)$ denotes the conditional of the LPS $\lambda_i(t_i)$ on $S_j(h_i) \times T_j$. In particular, if $s_i$ is rational for $t_i$ at $h_i$ then $s_i$ is rational with respect to the preference relation $\hat{P}_i$ and the set-valued belief $\hat{B}_{ij}(t_i, h_i)$, as defined above, where $\hat{P}_i$ is the preference relation on terminal nodes induced by $v_i(t_i)$.

Asheim's sufficient condition for backward induction is based on the notion of *admissible subgame consistency.* A type $t_i$ in an Asheim-model is said to be *admissible subgame consistent* with respect to a given profile $(\tilde{v}_i)_{i \in I}$ of utility functions if (1) $v_i(t_i) = \tilde{v}_i$, and (2) for every $h_i \in H_i^*$, the probability distribution $\lambda_i^{k_i(t_i,h_i)}(t_i)$ only assigns positive probability to strategy-type pairs $(s_j, t_j)$ such that $s_j$ is rational for $t_j$ at all $h_j \in H_j$ that follow $h_i$. In terms of our base model, this condition can be translated as: (1') $P_i(t_i) = \tilde{P}_i$, and (2') $t_i$ always believes in rationality at all future information sets. In fact, condition (2') is weaker than condition (2) since the notion of rationality in (2') is weaker than the notion of rationality in (2), but condition (2') would have sufficed to prove Asheim's theorem on backward induction.

In Proposition 7, Asheim shows that if a type $t_i$ respects common certain belief in the event that types are admissible subgame consistent with respect to $(\tilde{v}_i)_{i \in I}$, then $t_i$ has a unique strategy that is rational at all information sets, namely his backward induction strategy with respect to $(\tilde{v}_i)_{i \in I}$. Here, "certain belief in an event $E$" means that type $t_i$, in each of his probability distributions $\lambda_i^k(t_i)$, only assigns positive probability to types in $E$. In terms of our base model, this means that the type believes the event $E$ at each of his information sets. In Proposition

8, Asheim shows that common certain belief in admissible subgame consistency is possible. Translated in terms of our base model, Asheim's sufficient condition for backward induction may thus be written as follows:

**Asheim's condition BR:** Type $t_i$ respects common belief in the events that (1) types hold preference relations as specified by $(\tilde{P}_i)_{i \in I}$, and (2) types always believe in rationality at all future information sets.

### 3.2. Asheim & Perea's Model

Asheim and Perea (2005) propose a type-based semantic model that is very similar to Asheim's. Attention is restricted to two-player games, and an Asheim-Perea-model corresponds to a tuple

$$\mathcal{M} = ((T_i)_{i \in I}, (\lambda_i)_{i \in I}, (\ell_i)_{i \in I}, (v_i)_{i \in I}),$$

where $T_i$, $v_i$ and $\lambda_i$ are as in Asheim's model, and $\ell_i$ is a function that to every type $t_i$ and event $E \subseteq S_j \times T_j$ assigns some number $\ell_i(t_i, E) \in \{1, ..., K_i(t_i)\}$. (Recall that $K_i(t_i)$ denotes the number of probability distributions in $\lambda_i(t_i)$). The interpretation of $\ell_i$ is that $\ell_i(t_i, E)$ specifies the number of probability distributions in $\lambda_i(t_i)$ that are to be used in order to derive the conditional LPS of $\lambda_i(t_i)$ on $E$. The function $\ell_i$, however, is not relevant for our purpose here, and hence we will not elaborate more on it.

The sufficient condition for backward induction is based on the event that *types induce for every opponent's type a sequentially rational behavior strategy.* Consider a type $t_i$, and let $T_j^{t_i}$ be the set of types to which the LPS $\lambda_i(t_i)$ assigns positive probability (in some of its probability distributions). Asheim and Perea assume that for every $t_j \in T_j^{t_i}$ and every $s_j \in S_j$, the LPS $\lambda_i(t_i)$ assigns positive probability to $(s_j, t_j)$. For every information set $h_j \in H_j$ and action $a \in A(h_j)$, let $S_j(h_j, a)$ be the set of strategies in $S_j(h_j)$ that select action $a$ at $h_j$. Define for every type $t_j \in T_j^{t_i}$, $h_j \in H_j$ and $a \in A(h_j)$

$$\sigma_j^{t_i|t_j}(h_j)(a) := \frac{\lambda_i^k(t_i)(S_j(h_j, a) \times \{t_j\})}{\lambda_i^k(t_i)(S_j(h_j) \times \{t_j\})},$$

where $k$ is the first number such that $\lambda_i^k(t_i)(S_j(h_j) \times \{t_j\}) > 0$. The vector

$$\sigma_j^{t_i|t_j} = (\sigma_j^{t_i|t_j}(h_j)(a))_{h_j \in H_j, a \in A(h_j)}$$

is called the *behavior strategy induced by* $t_i$ for $t_j$. My interpretation of $\sigma_j^{t_i|t_j}(h_j)(a)$ is that type $t_i$ believes at every information set $h_i$ that type $t_j$ at information $h_j$ chooses action $a$ with probability $\sigma_j^{t_i|t_j}(h_j)(a)$, *unless* $h_j$ comes before $h_i$. If $h_j$ comes before $h_i$, namely, then there is a unique action $a^*$ at $h_j$ that leads to $h_i$, and hence $t_i$ *must* believe at $h_i$ that $t_j$ has chosen $a^*$ at $h_j$, whereas $\sigma_j^{t_i|t_j}(h_j)(a^*)$ may be less than one (in fact, may be zero). In all other cases, the

information that the game has reached $h_i$ does not give type $t_i$ additional information about the action choice of $t_j$ at $h_j$, and hence $\sigma_j^{t_i|t_j}(h_j)$ provides an intuitive candidate for the conditional belief of $t_i$ at $h_i$ about $t_j$'s behavior at $h_j$.

For every information set $h_j \in H_j$, let $\sigma_j^{t_i|t_j}|_{h_j}$ be the behavioral strategy that assigns probability one to all player $j$ actions preceding $h_j$, and coincides with $\sigma_j^{t_i|t_j}$ otherwise. The induced behavior strategy $\sigma_j^{t_i|t_j}$ is said to be *sequentially rational for* $t_j$ if at every information set $h_j \in H_j$, the behavior strategy $\sigma_j^{t_i|t_j}|_{h_j}$ only assigns positive probability to strategies in $S_j(h_j)$ that are rational for $t_j$ at $h_j$ (in the sense of Asheim's model above). Type $t_i$ is said to induce for every opponent's type a sequentially rational behavior strategy if for every $t_j \in T_j^{t_i}$ it is true that $\sigma_j^{t_i|t_j}$ is sequentially rational for $t_j$. As we have seen above, $\sigma_j^{t_i|t_j}$ represents for every $h_i \in H_i^*$ type $t_i$'s conditional belief at $h_i$ about player $j$'s behavior at future and parallel information sets. The requirement that $\sigma_j^{t_i|t_j}$ always be sequentially rational for $t_j$ thus means that $t_i$ always believes in rationality at all future and parallel information sets.

In Proposition 11, Asheim and Perea show that if a type $t_i$ respects common certain belief in the events that (1) types have utility functions as specified by $(\tilde{v}_i)_{i \in I}$, and (2) types induce for every opponent's type a sequentially rational behavior strategy, then $t_i$ has a unique strategy that is rational at all information sets, namely his backward induction strategy with respect to $(\tilde{v}_i)_{i \in I}$. The existence of such types follows from their Proposition 4 and the existence of a sequential equilibrium. In terms of our base model, Asheim and Perea's sufficient condition may thus be stated as follows:

**Asheim & Perea's condition BR:** Type $t_i$ respects common belief in the events that (1) types hold preference relations as specified by $(\tilde{P}_i)_{i \in I}$, and (2) types always believe in rationality at all future and parallel information sets.

### 3.3. Aumann's Model

Aumann (1995) proposes a state-based semantic model for extensive form structures with perfect information. An Aumann-model is a tuple

$$\mathcal{M} = (\Omega, (B_i)_{i \in I}, (f_i)_{i \in I}, (v_i)_{i \in I})$$

where $\Omega$ represents the set of states of the world, $B_i$ is a function that assigns to every state $\omega \in \Omega$ some subset $B_i(\omega)$ of states, $f_i$ is a function that assigns to every state $\omega$ some strategy $f_i(\omega) \in S_i$, and $v_i$ is a function that assigns to every $\omega$ some vNM utility function $v_i(\omega)$ on the set of terminal nodes. The functions $B_i$ must have the property that $\omega \in B_i(\omega)$ for all $\omega$, and for all $\omega, \omega' \in \Omega$ it must hold that $B_i(\omega)$ and $B_i(\omega')$ are either identical, or have an empty intersection. Hence, the set $\{B_i(\omega) | \omega \in \Omega\}$ is a partition of $\Omega$. The interpretation is that at state $\omega$, player $i$ believes that the true state is in $B_i(\omega)$. Aumann uses the term "knows" rather than "believes", but for the sake of uniformity we stick to the notion of belief here. The functions $f_i$ and $v_i$

must be measurable with respect to $B_i$, meaning that $f_i(\omega') = f_i(\omega)$ whenever $\omega' \in B_i(\omega)$, and similarly for $v_i$. The reason is that player $i$ cannot distinguish between states $\omega$ and $\omega'$, and hence his choice and preferences must be the same at both states.

It is problematic, however, to formally translate this model into conditional beliefs of our base model. Consider, for instance, a game with three players, in which players 1, 2 and 3 sequentially choose between *Stay* and *Leave*, and where *Leave* terminates the game. Consider a state $\omega$ where $f_1(\omega) = Leave$ and $B_2(\omega) = \{\omega\}$. Then, at player 2's information set, player 2 must conclude that the state cannot be $\omega$, but must be some state $\omega'$ with $f_1(\omega') = Stay$. However, there may be many such states $\omega'$, and hence it is not clear how player 2 should revise his belief about the state at his information set. Since his revised belief about the state will determine his revised belief about player 3's choice, it is not clear how to define player 2's revised belief about player 3's choice from Aumann's model.

It therefore seems necessary to carry out a somewhat informal translation into our base model, based on an interpretation of Aumann's main ideas. Aumann's model is essentially a static model, since for very state $\omega$ and every player $i$, his belief $B_i(\omega)$ is only defined at a single moment in time, presumably at the beginning of the game. At the same time, the static nature of the model suggests that players, upon observing that one of their information sets has been reached, do not revise more than "strictly necessary". In fact, the only beliefs that *must* be revised by player $i$ when finding out that his information set $h_i$ has been reached are, possibly, his beliefs about the opponents' choices at information sets preceding $h_i$. That is, if player 2 in the example above finds out that player 1 has chosen *Stay*, then this should not be a reason to change his belief about player 3's choice. Even stronger, player 2 *only* changes his belief about player 1's choice, while maintaining all his other beliefs, including his beliefs about the opponents' beliefs. That is, if we translate the nature of Aumann's model into our base model, then every type is supposed to never revise his belief about the opponents' choices, nor about the opponents' beliefs at future and parallel information sets. A type $t_i$, when arriving at some information set $h_i$, may only revise his belief about the opponents' *choices* at information sets that precede $h_i$ (but not about their types). For further reference, we call this condition the "no substantial belief revision condition".

The sufficient condition for backward induction presented by Aumann is *common knowledge of rationality*. Let $\omega$ be a state, $i$ a player and $h_i$ an information set controlled by $i$. At state $\omega$, player $i$ is said to be rational at information set $h_i$ if there is no $s_i \in S_i$ such that for every $\omega' \in B_i(\omega)$ it holds that

$$v_i(\omega)(z(s_i, (f_j(\omega'))_{j \neq i} | h_i)) > v_i(\omega)(z(f_i(\omega), (f_j(\omega'))_{j \neq i} | h_i)),$$

where $z(s_i, (f_j(\omega'))_{j \neq i} | h_i)$ is the terminal node that is reached if the game would start at $h_i$, and the players would choose in accordance with $(s_i, (f_j(\omega'))_{j \neq i})$. In terms of our base model, this means that strategy $f_i(\omega)$ is rational for player $i$ at $h_i$ with respect to the utility function $v_i(\omega)$ and his first-order belief $\{(f_j(\omega')_{j \neq i} \mid \omega' \in B_i(\omega)\}$ about the opponents' choices after $h_i$.

Let $\Omega^{rat}$ be the set of states $\omega$ such that at $\omega$ all players are rational at each of their information sets.

Common knowledge of rationality can now be defined by the following recursive procedure:

$$\begin{aligned} CKR^1 &= \Omega^{rat}; \\ CKR^k &= \{\omega \in CKR^{k-1} \mid B_i(\omega) \subseteq CKR^{k-1} \text{ for all players } i\} \end{aligned}$$

for $k \geq 2$. Then, common knowledge of rationality is said to hold at $\omega$ if $\omega \in CKR^k$ for all $k$. In Theorem A, Aumann proves that for every profile $(\tilde{v}_i)_{i \in I}$ of utility functions, for every state $\omega$ at which common knowledge of $(\tilde{v}_i)_{i \in I}$ and common knowledge of rationality hold, and for every player $i$, the strategy $f_i(\omega)$ is the backward induction strategy for player $i$ with respect to $(\tilde{v}_i)_{i \in I}$. In Theorem B, Aumann proves that there is an Aumann-model and a state $\omega$ at which common knowledge of $(\tilde{v}_i)_{i \in I}$ and common knowledge of rationality hold.

Since the beliefs $B_i(\omega)$ in Aumann's model correspond to initial belief in our base model, common knowledge of rationality corresponds to common initial belief in rationality at all information sets in our base model. By the latter we mean that a type (1) initially believes that all players choose rationally at all information sets, (2) initially believes that every type initially believes that all players choose rationally at all information sets, and so on. Together with the "no substantial belief revision condition" above, this implies that a type *always* believes that types initially believe that all players choose rationally at all information sets, and that a type always believes that types always believe that types initially believe that players choose rationally at all information sets, and so on. That is, Aumann's condition of common knowledge of rationality, together with the "no substantial belief revision condition", leads in our base model to common belief in the event that players initially believe in rationality at all information sets. Similarly, common knowledge of $(\tilde{v}_i)_{i \in I}$, together with the "no substantial belief revision condition", leads to common belief in the event that types have preferences according to $(\tilde{P}_i)_{i \in I}$, where $\tilde{P}_i$ is the preference relation that corresponds to $\tilde{v}_i$. That is, Aumann's sufficient conditions for backward induction may be translated into our base model as follows:

**Aumann's condition BR:** Type $t_i$ respects common belief in the events that (1) types hold preferences as specified by $(\tilde{P}_i)_{i \in I}$, (2) types initially believe in rationality at all information sets, and (3) types never revise their beliefs about the opponents' choices and beliefs at future and parallel information sets.

Clausing (2003) basically provides a reformulation of Aumann's model and definitions in a syntactic framework. Clausing's sufficient condition for backward induction is a little weaker than Aumann's, since Clausing only requires "true $(k-1)$th level belief" in rationality at all information sets, where $k$ is the maximal length of a path in the game tree, which is weaker than common knowledge of rationality as defined by Aumann. Quesada (2003) proves, in his Propositions 3.3 and 3.4, that Aumann's backward induction theorem can also be shown by taking Aumann's model and weakening some conditions imposed on the knowledge operators. However, since both models are identical in spirit to Aumann's, we omit a formal discussion of Clausing (2003) and Quesada (2003) in this overview.

### 3.4. Balkenborg & Winter's Model

Balkenborg and Winter (1997) present a state-based semantic model that is almost identical to Aumann's model, so we do not repeat it here. The only difference is that Balkenborg and Winter restrict attention to extensive form structures in which every player controls only one information set. In particular, the Balkenborg-Winter model also implies the "no substantial belief revision condition", as defined above, when translating it into terms of our base model. However, the sufficient conditions given for backward induction are different from Aumann's conditions, as they are based on the notion of *forward knowledge of rationality* rather than common knowledge of rationality.

For every player $i$, let $h_i$ be the unique information set controlled by player $i$. The definition of player $i$ being rational at $h_i$ is the same as in Aumann's model. Let $\Omega_i^{rat}$ be the set of states $\omega$ such that at $\omega$, player $i$ is rational at $h_i$. We say that player $j$ comes after player $i$ if $h_j$ comes after $h_i$. Forward knowledge of rationality can now be defined by the following recursive procedure. For every player $i$ define:

$$\begin{aligned} FKR_i^1 &= \Omega_i^{rat}; \\ FKR_i^k &= \{\omega \in FKR_i^{k-1} \mid B_i(\omega) \subseteq FKR_j^{k-1} \text{ for all } j \text{ that come after } i\}, \end{aligned}$$

for every $k \geq 2$. Then, forward knowledge of rationality is said to hold at state $\omega$ if $\omega \in FKR_i^k$ for all $i$ and all $k$. In Theorem 2.1, Balkenborg and Winter prove that for every profile $(\tilde{v}_i)_{i \in I}$ of utility functions, for every state $\omega$ at which common knowledge of $(\tilde{v}_i)_{i \in I}$ and forward knowledge of rationality hold, and for every player $i$, the strategy $f_i(\omega)$ is the backward induction strategy for player $i$ with respect to $(\tilde{v}_i)_{i \in I}$.

Since Balkenborg and Winter's model is essentially a static model, the belief $B_i(\omega)$ of player $i$ at $\omega$ must be interpreted as his initial belief, when reasoning in terms of our base model. If we assume that every player controls one information set, the notion of forward knowledge of rationality may thus be translated as follows into our base model: (1) type $t_i \in T_i$ initially believes that every player $j$ coming after $i$ chooses rationally at $h_j$, (2) $t_i$ initially believes that every player $j$ coming after $i$ believes initially that every player $k$ coming after $j$ chooses rationally at $h_k$, and so on. Together with the "no substantial belief revision condition", these conditions lead to the following event: (1) $t_i$ always believes that every player $j$ coming after $i$ chooses rationally at $h_j$, (2) $t_i$ always believes that every player $j$ coming after $i$ always believes that every player $k$ coming after $j$ chooses rationally at $h_k$, and so on. However, if one extends this notion to the case where players may control more than one information set, then one obtains the definition of forward belief in substantive rationality as given in Definition 2.6. As such, Balkenborg and Winter's condition of forward knowledge of rationality, together with the "no substantial belief revision condition", corresponds to forward belief in substantive rationality in our base model. Balkenborg and Winter's sufficient condition for backward induction, phrased in terms of our base model, is thus as follows:

**Balkenborg & Winter's condition BR:** Type $t_i$ (1) respects common belief in the event that types hold preferences as specified by $(\tilde{P}_i)_{i\in I}$, (2) respects forward belief in substantive rationality, and (3) respects common belief in the event that types never revise their beliefs about the opponents' choices and beliefs at future and parallel information sets.

Quesada (2003) proves, in his Proposition 3.1, that Balkenborg and Winter's sufficient condition for backward induction would still be sufficient if some conditions on the knowledge operators would be weakened.

### 3.5. Clausing's Model

Clausing (2004) presents a syntactic model for games with perfect information. For our purposes here it is not necessary to discuss the complete formalism of Clausing's model, and therefore we restrict ourselves to presenting only the key ingredients. A Clausing-model is a tuple

$$\mathcal{M} = (L, (\hat{B}_i)_{i\in I}, (v_i)_{i\in I})$$

where $L$ is a language, or set of statements, $\hat{B}_i$ is a function that assigns to every statement $f \in L$ some subset $\hat{B}_i(f) \subseteq L$ of statements, and $v_i$ is a utility function for player $i$ on the set of terminal nodes. By "$g \in \hat{B}_i(f)$" we mean the statement that "player $i$ believes statement $g$ upon learning that $f$ holds". It is assumed that $L$ contains all statements of the form "player $i$ chooses strategy $s_i$", and that it is closed under the operations $\neg$ (not), $\wedge$ (and) and $\hat{B}_i$. By the latter, we mean that if $f$ and $g$ are statements in $L$, then so are the statements "$\neg f$", "$f \wedge g$" and "$g \in \hat{B}_i(f)$".

Clausing's sufficient condition for backward induction is *forward belief from the root to all information sets $h$ in rationality at $h$*. We say that strategy $s_i$ is rational for player $i$ at information set $h_i$ if there is no other strategy $s_i' \in S_i(h_i)$ such that player $i$ would believe, upon learning that $h_i$ has been reached, that $s_i'$ would lead to a higher utility than $s_i$. Formally, there should be no $s_i' \in S_i(h_i)$ and no statement $f \in L$ about the opponents' strategy choices such that (1) player $i$ believes $f$ upon learning that all opponents $j$ have chosen a strategy in $S_j(h_i)$, and (2) for every opponents' strategy profile $s_{-i}$ compatible with $f$ it would be true that $v_i(z(s_i', s_{-i}|h_i)) > v_i(z(s_i, s_{-i}|h_i))$. Player $i$ is said to believe at $h_i$ that player $j$ is rational at $h_j$ if, upon learning that $h_i$ has been reached, player $i$ believes the statement "player $j$ chooses a strategy that is rational for $j$ at $h_j$". Forward belief from the root to all information sets $h$ in rationality at $h$ can now be defined by the following sequence of statements:

$$FB_i^1(h_i) \;=\; \text{"player } i \text{ believes, upon learning that } h_i \text{ has been reached, that every opponent } j \text{ will be rational at all } h_j \text{ that follow } h_i\text{"}$$

for all players $i$ and all $h_i \in H_i^*$, and

$$FB_i^k(h_i) \;=\; \text{"player } i \text{ believes, upon learning that } h_i \text{ has been reached, the statement } FB_j^{k-1}(h_j) \text{ for all opponents } j \text{ and all } h_j \text{ that follow } h_i\text{"}$$

for all players $i$, $h_i \in H_i^*$ and $k \geq 2$. Player $i$ is said to respect forward belief from the root to all information sets $h$ in rationality at $h$ if for every $h_i$, player $i$ believes, upon learning that $h_i$ has been reached, the statements $FB_j^k(h_j)$ for all $k$, all opponents $j$ and all $h_j \in H_j$ that follow $h_i$. In Proposition 2, Clausing shows that this condition implies backward induction, whereas his Proposition 3 demonstrates that this condition is possible. In the language of our base model, Clausing's condition clearly corresponds to forward belief in substantive rationality.

**Clausing's condition BR:** Type $t_i$ (1) respects common belief in the event that types hold preferences as specified by $(\tilde{P}_i)_{i \in I}$, and (2) respects forward belief in substantive rationality.

### 3.6. Feinberg's Model

Feinberg (2005) provides a syntactic model for dynamic games, with or without perfect information, which is similar to Clausing's model. Since a full treatment of Feinberg's model would take us too far afield, we present a highly condensed version of his model here, which will serve for our restricted purposes. A Feinberg-model is a tuple

$$\mathcal{M} = (L, (C_i)_{i \in I}, (v_i)_{i \in I})$$

where $L$ is a language, or set of statements, $C_i$ is a function that selects for every information set $h_i \in H_i^*$ a set $C_i(h_i) \subseteq L$ of statements, and $v_i$ is a utility function for player $i$ on the set of terminal nodes. The interpretation of $f \in C_i(h_i)$ is that player $i$ is *confident* of statement $f$ at information set $h_i$. The language $L$ must contain all statements of the form "player $i$ chooses strategy $s_i$", and must be closed under the application of the operators $\neg$ (not), $\wedge$ (and) and $C_i(h_i)$. By the latter, we mean that, if $f$ is a statement in $L$, then the statement "$f \in C_i(h_i)$" must also be in $L$.

Feinberg presents two different sufficient conditions for backward induction, namely *common confidence of hypothetical rationality* and *iterated future confidence of rationality*. Strategy $s_i$ is said to be rational for player $i$ at $h_i$ if there is no other strategy $s_i' \in S_i(h_i)$ such that player $i$ would be confident at $h_i$ that $s_i'$ would lead to a higher utility than $s_i$. By the latter, we mean that there should be no $s_i' \in S_i(h_i)$, and no statement $f$ about the opponents' strategy choices, such that (1) $i$ is confident of $f$ at $h_i$, and (2) for every opponents' strategy profile $s_{-i}$ compatible with $f$ it would hold that $v_i(z(s_i', s_{-i})|h_i) > v_i(z(s_i, s_{-i})|h_i)$. We say that $i$ is confident at $h_i$ that $j$ is rational at $h_j$ if the statement "player $j$ chooses a strategy that is rational for $j$ at $h_j$" belongs to $C_i(h_i)$. Common confidence in hypothetical rationality can now be defined recursively be the following sequence of statements:

$$CCHR^1 = \text{"every player } i \text{ is confident at every } h_i \text{ that every opponent } j \text{ will be rational at every } h_j \text{ not preceding } h_i\text{"}$$

and, for every $k \geq 2$,

$$CCHR^k = \text{"every player } i \text{ is confident at every } h_i \text{ of } CCHR^{k-1}\text{"}.$$

Player $i$ is said to respect common confidence in hypothetical rationality if, for every $h_i$ and every $k$, player $i$ is confident at $h_i$ of $CCHR^k$. In Proposition 10, Feinberg shows that this condition is possible, and implies backward induction. In terms of our base model, this condition corresponds exactly to our definition of common belief in the event that types always believe in rationality at all future and parallel information sets.

**Feinberg's first condition BR:** Type $t_i$ respects common belief in the events that (1) types hold preference relations as specified by $(\tilde{P}_i)_{i \in I}$, and (2) types always believe in rationality at all future and parallel information sets.

Iterated future confidence of rationality can be defined by means of the following sequence of statements:

$$IFCR_i^1(h_i) \quad = \quad \text{"player } i \text{ is confident at } h_i \text{ that all opponents } j \text{ will be rational at all } h_j \text{ that follow } h_i\text{"}$$

for all $i \in I$ and all $h_i \in H_i^*$, and

$$IFCR_i^k(h_i) \quad = \quad \text{"player } i \text{ is confident at } h_i \text{ of } IFCR_j^{k-1}(h_j) \text{ for all opponents } j \text{ and all } h_j \text{ that follow } h_i\text{"}$$

for all $i \in I$, $h_i \in H_i^*$ and $k \geq 2$. Player $i$ is said to respect iterated future confidence of rationality if, for every $k$, every $h_i$, every opponent $j$, and every $h_j$ following $h_i$, player $i$ is confident at $h_i$ of $IFCR_j^k(h_j)$. Feinberg shows in his Proposition 11 that this condition is possible and leads to backward induction. Translated into our base model, this condition corresponds to our definition of forward belief in substantive rationality.

**Feinberg's second condition BR:** Type $t_i$ (1) respects common belief in the event that types hold preferences as specified by $(\tilde{P}_i)_{i \in I}$, and (2) respects forward belief in substantive rationality.

### 3.7. Perea's Model

Perea (2005) proposes a type-based semantic model that is very similar to our base model. The difference is that in Perea (2005) the players' initial and revised beliefs are assumed to be point-beliefs, that is, contain exactly one strategy-type pair for each opponent. Moreover, in Perea (2005) the model is assumed to be *complete,* which will be defined below. A Perea-model is a tuple

$$\mathcal{M} = ((T_i)_{i \in I}, (\hat{B}_{ij})_{j \neq i}, (P_i)_{i \in I})$$

where $T_i$ is player $i$'s set of types, $P_i$ assigns to every type $t_i \in T_i$ a strict preference relation $P_i(t_i)$ over the terminal nodes, $\hat{B}_{ij}$ assigns to every type $t_i \in T_i$ and every information set $h_i \in H_i^*$ a belief $\hat{B}_{ij}(t_i, h_i) \subseteq S_j(h_i) \times T_j$ consisting of only *one* strategy-type pair, and the

model $\mathcal{M}$ is *complete*. By complete, we mean that for every player $i$, every strict preference relation $\hat{P}_i$ and every belief vector $(\tilde{B}_{ij})_{j \neq i}$, assigning to every opponent $j$ and $h_i$ some point belief $\tilde{B}_{ij}(h_i) \in S_j(h_i) \times T_j$, there is some type $t_i \in T_i$ with $P_i(t_i) = \hat{P}_i$ and $\hat{B}_{ij}(t_i, h_i) = \tilde{B}_{ij}(h_i)$ for all $j$ and $h_i$.

Perea's sufficient condition for backward induction is common belief in the events that (1) players initially believe in $(\tilde{P}_i)_{i \in I}$, (2) players initially believe in rationality at all information sets, and (3) the players' belief revision procedures satisfy some form of minimal belief revision. The crucial difference with the other models discussed here is that condition (1) allows players to revise their belief about the opponents' preference relations as the game proceeds. On the other hand, conditions (2) and (3) together imply that players should *always* believe that every opponent chooses rationally at *all information sets;* a condition that cannot be realized in general if players do not revise their beliefs about the opponents' preference relations.

A type $t_i$ is said to initially believe in $(\tilde{P}_i)_{i \in I}$ if for every opponent $j$, the initial belief $\hat{B}_{ij}(t_i, h_0)$ consists of a strategy-type pair $(s_j, t_j)$ where $P_j(t_j) = \tilde{P}_j$. In order to formalize condition (3), we need the definition of an elementary statement. A first-order elementary statement about player $i$ is a statement of the form "player $i$ has a certain preference relation" or "player $i$ believes at $h_i$ that opponent $j$ chooses a certain strategy". Recursively, one can define, for every $k \geq 2$, a $k$-th order elementary statement about player $i$ as a statement of the form "player $i$ believes at $h_i$ that $\varphi$" where $\varphi$ is a $(k-1)$-th order elementary statement. An elementary statement about player $i$ is then an elementary statement about player $i$ of some order $k$. Now, let $h_i \in H_i \backslash h_0$, and let $h_i'$ be the information set in $H_i^*$ that precedes $h_i$ and for which no other player $i$ information set is between $h_i'$ and $h_i$. For every opponent $j$, let $(s_j', t_j')$ be the strategy-type pair in $\hat{B}_{ij}(t_i, h_i')$, and let $(s_j, t_j)$ be the strategy-type pair in $\hat{B}_{ij}(t_i, h_i)$. Type $t_i$ is said to satisfy minimal belief revision at $h_i$ if for every opponent $j$ the strategy-type pair $(s_j, t_j)$ is such that (1) $s_j$ is rational for $t_j$ at all information sets, (2) there is no other strategy type pair $(s_j'', t_j'')$ in $S_j(h_i) \times T_j$ satisfying (1) such that $t_j''$ and $t_j'$ disagree on fewer elementary statements about player $j$ than $t_j$ and $t_j'$ do, and (3) there is no other strategy-type pair $(s_j'', t_j'')$ in $S_j(h_i) \times T_j$ satisfying (1) and (2) such $P_j(t_j'')$ and $P_j(t_j')$ disagree on fewer pairwise rankings of terminal nodes than $P_j(t_j)$ and $P_j(t_j')$ do. In particular, minimal belief revision requires that a type always believes that his opponents choose rationally at *all* information sets. Note that for the definition of minimal belief revision it is very important that the model $\mathcal{M}$ is assumed to complete. Theorem 5.1 in Perea (2005) shows that there is a Perea-model which satisfies the sufficient condition listed above. Theorem 5.2 in that paper demonstrates that this sufficient condition leads to backward induction. As such, Perea's sufficient condition for backward induction can be stated as follows in terms of our base model:

**Perea's condition BR:** Type $t_i$ respects common belief in the events that (1) types hold point-beliefs, (2) types initially believe in $(\tilde{P}_i)_{i \in I}$, (3) types always believe in rationality at all information sets, and (4) types satisfy minimal belief revision.

### 3.8. Quesada's Model

Quesada (2002) presents a model for games with perfect information which is neither semantic nor syntactic. The key ingredient is to model the players' uncertainty by means of *Bonanno-belief systems* (Bonanno (1992)). A Bonnano-belief system is a profile $\beta = (\beta_i)_{i \in I}$, where $\beta_i$ is a belief vector that assigns to every information set $h$ (not necessarily controlled by player $i$) some terminal node $\beta_i(h)$ which follows $h$. The interpretation is that player $i$, upon learning that the game has reached information set $h$, believes that he and his opponents will act in such a way that terminal node $\beta_i(h)$ will be reached. A Quesada-model is a pair

$$\mathcal{M} = (\mathcal{B}, (v_i)_{i \in I})$$

where $\mathcal{B}$ is a set of Bonnano-belief systems, and $v_i$ is a utility function for player $i$ over the terminal nodes. Quesada's sufficient condition for backward induction states that every belief system in $\mathcal{B}$ should be *rational*, and that every belief system in $\mathcal{B}$ should be *justifiable* by other belief systems in $\mathcal{B}$. Formally, a belief system $\beta = (\beta_i)_{i \in I}$ is said to be rational if for every player $i$ and every information set $h_i \in H_i$ it holds that $v_i(\beta_i(h_i)) \geq v_i(\beta_i((h_i, a)))$ for every action $a \in A(h_i)$, where $(h_i, a)$ denotes the information set that immediately follows action $a$ at $h_i$. We say that belief system $\beta = (\beta_i)_{i \in I}$ in $\mathcal{B}$ is justifiable by other belief systems in $\mathcal{B}$ if for every player $i$, every $h_i \in H_i$, every opponent $j$, and every $h_j \in H_j$ between $h_i$ and $\beta_i(h_i)$ there is some belief system $\beta' = (\beta'_i)_{i \in I}$ in $\mathcal{B}$ such that $\beta'_j(h_j) = \beta_i(h_i)$. A belief system $\beta = (\beta_i)_{i \in I}$ is called the backward induction belief system if for every player $i$ and every information set $h$, $\beta_i(h)$ is the terminal node which is reached by applying the backward induction procedure (with respect to $(v_i)_{i \in I}$) from $h$ onwards. In Proposition 1, Quesada shows that there is one, and only one, set $\mathcal{B}$ which satisfies the two conditions above, namely the set containing only the backward induction belief system.

We shall now translate these conditions into our base model. Take a set $\mathcal{B}$ of belief systems such every belief system in $\mathcal{B}$ is justifiable by other belief systems in $\mathcal{B}$ (and thus satisfies Quesada's second condition above). Then, every belief vector $\beta_i$ in $\mathcal{B}$ induces, for every $h_i$, a point-belief about the opponents' strategy choices as follows: For every $h_i$ there is some opponents' strategy profile $s_{-i}(\beta_i, h_i) \in S_{-i}(h_i)$ such that, for every action $a \in A(h_i)$, the action $a$ followed by $s_{-i}(\beta_i, h_i)$ leads to the terminal node $\beta_i(h_i, a)$. Hence, $s_{-i}(\beta_i, h_i)$ may be interpreted as $\beta_i$'s conditional point-belief at $h_i$ about the opponents' strategy choices. (Note that this belief need not be unique, as $\beta_i$ does not restrict player $i$'s beliefs at $h_i$ about opponents' choices at parallel information sets). The belief vector $\beta_i$ also induces, for every $h_i$, a conditional point-belief about the opponents' belief vectors $\beta'_j$ in $\mathcal{B}$. Consider, namely, an information set $h_i \in H_i$, some opponent $j$ and an information set $h_j$ between $h_i$ and the terminal node $\beta_i(h_i)$ such that there is no further player $j$ information set between $h_i$ and $h_j$. Since $\mathcal{B}$ satisfies Quesada's justifiability condition, there is some player $j$ belief vector $\beta_j(\beta_i, h_i)$ in $\mathcal{B}$ such that $\beta_j(\beta_i, h_i)(h_j) = \beta_i(h_i)$. (Again, this choice need not be unique). This belief vector $\beta_j(\beta_i, h_i)$ may then serve as $\beta_i$'s conditional point-belief at $h_i$ about player $j$'s belief vector. Summarizing, every

belief vector $\beta_i$ induces, at every $h_i$, a conditional point-belief about the opponents' strategy choices and the opponents' belief vectors.

Now, if we interpret every belief vector $\beta_i$ in $\mathcal{B}$ as a type $t_i(\beta_i)$ in our base model, then, by the insights above, every type $t_i(\beta_i)$ induces, at every $h_i$, a conditional point-belief about the opponents' strategy choices and types $t_j(\beta_j)$. Hence, similarly to Perea's model, Quesada's model can be translated into our base model by imposing common belief in the event that types hold point-beliefs. Let $T_i(\mathcal{B})$ denote the set of all such types $t_i(\beta_i)$ induced by some belief vector $\beta_i$ in $\mathcal{B}$. A combination of Quesada's rationality condition and justifiability condition implies that, whenever $\beta_i$ in $\mathcal{B}$ believes at $h_i$ that player $j$ chooses action $a$ at some $h_j$ between $h_i$ and $\beta_i(h_i)$ (with no player $j$ information set between $h_i$ and $h_j$), then there is some rational belief vector $\beta_j(\beta_i, h_i)$ in $\mathcal{B}$ such that $\beta_j(\beta_i, h_i)(h_j) = \beta_i(h_i)$. In particular, action $a$ must be part of the rational belief vector $\beta_j(\beta_i, h_i)$, and hence action $a$ must be optimal with respect to $\beta_j(\beta_i, h_i)$. In terms of our base model, this means that, whenever type $t_i(\beta_i)$ believes at $h_i$ that information set $h_j$ will be reached in the future, and believes at $h_i$ that player $j$ will choose action $a$ at $h_j$, then $t_i(\beta_i)$ must believe at $h_i$ that player $j$ is of some type $t_j(\beta_j)$ for which $a$ is rational. In other words, every type $t_i(\beta_i)$ in $T_i(\mathcal{B})$ always believes in rationality at future information sets that are believed to be reached. However, since $t_i(\beta_i)$ believes at every information set that every opponent $j$ is of some type $t_j(\beta_j)$ in $T_j(\mathcal{B})$, it follows that every $t_i(\beta_i)$ in $T_i(\mathcal{B})$ always believes in the event that all types believe in rationality at future information sets that believed to be reached. By recursively applying this argument, one may conclude that every $t_i(\beta_i)$ in $T_i(\mathcal{B})$ respects common belief in the event that types always believe in rationality at future information sets that are believed to be reached. Quesada's sufficient condition can thus be formulated as follows in terms of our base model:

**Quesada's condition BR:** Type $t_i$ respects common belief in the events that (1) types hold preferences as specified by $(\tilde{P}_i)_{i \in I}$ , (2) types hold point-beliefs, and (3) types always believe in rationality at future information sets that are believed to be reached.

### 3.9. Samet's Model

Samet (1996) presents a state-based semantic model which is an extension of the models by Aumann (1995) and Balkenborg and Winter (1997). A Samet-model is a tuple

$$\mathcal{M} = (\Omega, (B_i)_{i \in I}, (f_i)_{i \in I}, (\tau_i)_{i \in I}, (v_i)_{i \in I}),$$

where $\Omega, B_i, f_i$ and $v_i$ are as in the Aumann-model, and $\tau_i$ is a so-called *hypothesis transformation* that assigns to every state $\omega$ and non-empty event $E \subseteq \Omega$ some new state $\omega'$. My interpretation of $\tau_i$ is that if player $i$ currently believes that the state is in $B_i(\omega)$, but later observes the event $E$, then he will believe that the state is in $B_i(\omega') \cap E$. Samet defines the hypothesis transformation in a different, but equivalent, way. In Samet's terminology, a hypothesis transformation assigns to every initial belief $B_i(\omega)$ and event $E$ some new belief $B_i(\omega')$ for some $\omega' \in \Omega$. However, this definition is equivalent to the existence of a function $\tau_i$ as described in our model. The function

$\tau_i$ must satisfy the following two conditions: (1) $B_i(\tau_i(\omega, E)) \cap E$ is nonempty for every $\omega$ and $E$, and (2) $\tau_i(\omega, E) = \omega$ whenever $B_i(\omega)$ has a nonempty intersection with $E$. These conditions indicate that $B_i(\tau_i(\omega, E)) \cap E$ may be interpreted as a well-defined conditional belief for player $i$ at state $\omega$ when observing the event $E$.

As to the functions $f_i$, mapping states to strategy choices, it is assumed that for every terminal node $z$ there is some state $\omega \in \Omega$ such that the profile $(f_i(\omega))_{i \in I}$ of stategies reaches $z$. This implies that for every information set $h_i$, the event

$$[h_i] = \{\omega \in \Omega \mid (f_i(\omega))_{i \in I} \text{ reaches } h_i\}$$

is nonempty, and hence can be used as conditioning event for the hypothesis transformation $\tau_i$. Samet assumes in his model a function $\xi$ (instead of $(f_i)_{i \in I}$) mapping states to terminal nodes, and assumes that for every terminal node $z$ there is some $\omega \in \Omega$ with $\xi(\omega) = z$. However, he shows that this function $\xi$ induces, in some precise way, a profile $(f_i)_{i \in I}$ of strategy functions, as we use it. We work directly with the strategy functions here, in order to make the model as similar as possible to the Aumann-model and the Balkenborg-Winter-model.

In contrast to Aumann's model and Balkenborg and Winter's model, every state $\omega$ in Samet's model formally induces a conditional belief vector in our base model. Namely, take some state $\omega$, a player $i$, and some information set $h_i \in H_i^*$. Then,

$$\hat{B}_i(\omega, h_i) := B_i(\tau_i(\omega, [h_i])) \cap [h_i]$$

respresents player $i$'s conditional belief at $h_i$ about the state. Since every state $\omega'$ induces for player $j$ a strategy choice $f_j(\omega')$ and a conditional belief vector $(\hat{B}_j(\omega', h_j))_{h_j \in H_j^*}$, first-order conditional beliefs about the opponents' strategies, and higher-order conditional beliefs about the opponents' conditional beliefs can be derived at every state with the help of the hypothesis transformations $\tau_i$. Hence, Samet's model can be directly and formally translated into our base model.

Samet's sufficient condition for backward induction is *common hypothesis of node rationality.* At state $\omega$, player $i$ said to be rational at $h_i \in H_i$ if (1) $\omega \in [h_i]$, and (2) there is no $s_i \in S_i$ such that for every $\omega' \in B_i(\omega) \cap [h_i]$ it holds that

$$v_i(\omega)(z(s_i, (f_j(\omega'))_{j \neq i} | h_i)) > v_i(\omega)(z(f_i(\omega), (f_j(\omega'))_{j \neq i} | h_i)),$$

where the definition of this expression is as in Aumann's model. Let $[rat_i(h_i)]$ denote the set of states $\omega$ such that at $\omega$, player $i$ is rational at $h_i$. Common hypothesis of node rationality can now be defined by the following recursive procedure: For every player $i$ and information set $h_i \in H_i^*$, let

$$CHNR(h_i, h_i) = [rat_i(h_i)].$$

Note that, by condition (1) above, $CHNR(h_i, h_i)$ only contains states at which $h_i$ is indeed reached. Now, let $k \geq 0$, and suppose that $CHNR(h_i, h_j)$ has been defined for all information

sets $h_i \in H_i^*, h_j \in H_j^*$ such that $h_j$ comes after $h_i$, and there are at most $k$ information sets between $h_i$ and $h_j$. Suppose now that $h_j$ comes after $h_i$, and that there are exactly $k+1$ information sets between $h_i$ and $h_j$. Let $h$ be the unique information set that immediately follows $h_i$ and precedes $h_j$. Define

$$CHNR(h_i, h_j) = \{\omega \in \Omega \mid B_i(\tau_i(\omega, [h])) \cap [h] \subseteq CHNR(h, h_j)\}.$$

Common hypothesis of node rationality is said to hold at state $\omega$ if $\omega \in CHNR(h_0, h)$ for all information sets $h$. Hence, the player at $h_0$ believes that (1) every opponent $j$ will choose rationally at those information sets $h_j$ that immediately follows $h_0$, and which he believes to be reached from $h_0$, (2) every such opponent $j$ will believe at every such $h_j$ that every other player $k$ will choose rationally at those $h_k$ that immediately follows $h_j$, and which he believes to be reached from $h_j$, and so on.

Samet shows in Theorem 5.3 that for every profile $(v_i)_{i \in I}$ of utility functions, for every state $\omega$ at which common knowledge of $(v_i)_{i \in I}$ and common hypothesis of node rationality hold, the strategy profile $(f_i(\omega))_{i \in I}$ leads to the backward induction outcome with respect to $(v_i)_{i \in I}$. In particular, the player at $h_0$ chooses the backward induction action at $h_0$ with respect to $(v_i)_{i \in I}$. In Theorem 5.4, Samet shows that there always exists some state $\omega$ at which common knowledge of $(v_i)_{i \in I}$ and common hypothesis of node rationality hold.

For a given state $\omega$ and information set $h_i \in H_i^*$, say that common hypothesis of node rationality *at* $h_i$ holds if $\omega \in CHNR(h_i, h)$ for all information sets $h$ that follow $h_i$. Then, Samet's Theorem 5.3 can be generalized as follows: For every $h_i \in H_i$ and every $\omega$ at which common knowledge of $(v_i)_{i \in I}$ and common hypothesis of node rationality at $h_i$ hold, the strategy $f_i(\omega)$ chooses at $h_i$ the backward induction action with respect to $h_i$.

In order to translate this sufficient condition into our base model, it is important to understand all implications of common hypothesis of node rationality. By definition, common hypothesis of node rationality at $h_i$ implies that player $i$ believes at $h_i$ that (1) every opponent $j$ will choose rationally at every information set $h_j$ that immediately follows $h_i$, (2) every such opponent $j$ will believe at every such $h_j$ that every other player $k$ will choose rationally at every $h_k$ that immediately follows $h_j$, and so on. However, there are more implications.

Consider namely an information set $h_j \in H_j$ that immediately follows $h_i$ and some information set $h_k \in H_k$ which immediately follows $h_j$ such that $B_i(\tau_i(\omega, [h_j])) \subseteq [h_k]$. Hence, in terms of our base model, player $i$ believes at $h_i$ that $h_k$ will be reached. Suppose that state $\omega$ is such that common hypothesis of node rationality at $h_i$ holds at $\omega$. By (1) above, it holds that (1') $B_i(\tau_i(\omega, [h_j])) \cap [h_j] \subseteq [rat_j(h_j)]$. By (2) above, it holds for every $\omega' \in B_i(\tau_i(\omega, [h_j])) \cap [h_j]$ that (2') $B_j(\tau_j(\omega', [h_k])) \cap [h_k] \subseteq [rat_k(h_k)]$. However, since $B_i(\tau_i(\omega, [h_j])) \subseteq [h_k]$, it follows that $\omega' \in [h_k]$ for every $\omega' \in B_i(\tau_i(\omega, [h_j]))$, and hence $B_j(\tau_j(\omega', [h_k])) \cap [h_k] = B_j(\omega')$ for every $\omega' \in B_i(\tau_i(\omega, [h_j])) \cap [h_j]$. By (2') it thus follows that $B_j(\omega') \subseteq [rat_k(h_k)]$ for every $\omega' \in B_i(\tau_i(\omega, [h_j])) \cap [h_j]$. Since $\omega' \in B_j(\omega')$, it follows in particular, $\omega' \in [rat_k(h_k)]$ for every $\omega' \in B_i(\tau_i(\omega, [h_j])) \cap [h_j]$, which means that player $i$ believes at $h_i$ that player $k$ chooses rationally at $h_k$. Hence, we have shown that common hypothesis of node rationality at $h_i$ implies

that player $i$ believes at $h_i$ that player $k$ chooses rationally at $h_k$ whenever (1) there is only one information set between $h_i$ and $h_k$, and (2) player $i$ believes at $h_i$ that $h_k$ will be reached. By induction, one can now show that common hypothesis of node rationality at $h_i$ implies that player $i$ believes at $h_i$ that player $k$ chooses rationally at $h_k$ whenever (1) $h_k$ follows $h_i$ and (2) player $i$ believes at $h_i$ that $h_k$ can be reached.

By a similar argument, one can show that common hypothesis of node rationality at $h_i$ implies that player $i$ believes at $h_i$ that common hypothesis of node rationality will hold at every future information set $h_j$ which player $i$ believes to be reached from $h_i$. Together with our previous insight, this means that common hypothesis of node rationality may be translated into our base model by forward belief in material rationality (see our Definition 2.7). Samet's sufficient condition for backward induction, phrased in terms of our base model, is thus as follows:

**Samet's condition BR:** Type $t_i$ (1) respects common belief in the event that types hold preferences as specified by $(\tilde{P}_i)_{i \in I}$, and (2) respects forward belief in material rationality.

### 3.10. Stalnaker's Model

Stalnaker (1998) proposes a state-based semantic model for perfect information games in which every information set is controlled by a different player. The model we present here is not an exact copy of Stalnaker's model, but captures its essential properties. A Stalnaker-model is a tuple

$$\mathcal{M} = (\Omega, (\lambda_i)_{i \in I}, (f_i)_{i \in I}, (v_i)_{i \in I})$$

where $\Omega, f_i$ and $v_i$ are as in the Aumann-model, and $\lambda_i$ is a function that assigns to every state $\omega$ some lexicographic probability system (see Asheim's model) $\lambda_i(\omega)$ on $\Omega$. That is, $\lambda_i(\omega)$ is a sequence $(\lambda_i^1(\omega), ..., \lambda_i^{K_i(\omega)}(\omega))$ where $\lambda_i^k(\omega)$ is a probability distribution on $\Omega$. For every information set $h$ let $[h] = \{\omega \in \Omega \mid (f_i(\omega))_{i \in I}$ reaches $h\}$. We assume that $[h]$ is non-empty for all $h$, and that $\lambda_i(\omega)$ has full support on $\Omega$. By the latter, we mean that for every $\omega \in \Omega$ there is some probability distribution $\lambda_i^k(\omega)$ in $\lambda_i(\omega)$ with $\lambda_i^k(\omega)(\omega) > 0$. As such, $\lambda_i$ and $(f_j)_{j \neq i}$ induce, for every state $\omega$, a probabilistic belief revision policy for player $i$ in the following way. For every $h_i \in H_i^*$, let $k_i(\omega, h_i)$ be the first $k$ such that $\lambda_i^k(\omega)$ assigns positive probability to $[h_i]$. Then, the probability distribution $\mu_i(\omega, h_i)$ on $[h_i]$ given by

$$\mu_i(\omega, h_i)(\omega') = \frac{\lambda_i^{k_i(\omega, h_i)}(\omega')}{\lambda_i^{k_i(\omega, h_i)}([h_i])}$$

for every $\omega' \in [h_i]$ represents player $i$'s revised belief at $\omega$ upon observing that $h_i$ has been reached. More generally, for every event $E \subseteq \Omega$, the probability distribution $\mu_i(\omega, E)$ on $E$ given by

$$\mu_i(\omega, E)(\omega') = \frac{\lambda_i^{k_i(\omega, E)}(\omega')}{\lambda_i^{k_i(\omega, E)}(E)}$$

for every $\omega' \in E$ defines player $i$'s revised belief upon receiving information $E$. Here, $k_i(\omega, E)$ is the first $k$ such that $\lambda_i^k(\omega)$ assigns positive probability to $E$. The LPS $\lambda_i(\omega)$ naturally induces, for every information set $h_i \in H_i^*$, the non-probabilistic conditional belief

$$\hat{B}_i(\omega, h_i) := \text{supp}\mu_i(\omega, h_i),$$

and hence Stalnaker's model can be translated directly into our base model.

Stalnaker's sufficient condition for backward induction consists of *common initial belief in sequential rationality, and common belief in the event that players treat information about different players as epistemically independent.* Player $i$ is called sequentially rational at $\omega$ if at every information set $h_i \in H_i^*$, the strategy $f_i(\omega)$ is optimal given the utility function $v_i(\omega)$ and the revised belief about the opponents' strategy choices induced by $\mu_i(\omega, h_i)$ and $(f_j)_{j \neq i}$. Let $\Omega^{srat}$ be the set of states at which all players are sequentially rational. Common initial belief in sequential rationality can be defined by the following recursive procedure:

$$\begin{aligned} CIBSR^1 &= \Omega^{srat}; \\ CIBSR^k &= \{\omega \in CIBSR^{k-1} \mid \hat{B}_i(\omega, h_0) \subseteq CIBSR^{k-1} \text{ for all players } i\} \end{aligned}$$

for all $k \geq 2$. Common initial belief in sequential rationality is said to hold at $\omega$ if $\omega \in CIBSR^k$ for all $k$. We say that two states $\omega$ and $\omega'$ are indistinguishable for player $i$ if $f_i(\omega) = f_i(\omega')$, $v_i(\omega) = v_i(\omega')$ and $\mu_i(\omega, h_i) = \mu_i(\omega', h_i)$ for all $h_i \in H_i^*$. An event $E$ is said to be about player $i$ if for every two states $\omega, \omega'$ that are indistinguishable for player $i$, either both $\omega$ and $\omega'$ are in $E$, or none is in $E$. We say that at $\omega$ player $i$ treats information about different players as epistemically independent if for every two different opponents $j$ and $k$, for every event $E_j$ about player $j$ and every event $E_k$ about player $k$, it holds that $\mu_i(\omega, E_j)(E_k) = \mu_i(\omega, \Omega \backslash E_j)(E_k)$ and $\mu_i(\omega, E_k)(E_j) = \mu_i(\omega, \Omega \backslash E_k)(E_j)$. In his theorem on page 43, Stalnaker shows that common initial belief in sequential rationality and common belief in the event that players treat information about different players as epistemically independent leads to backward induction.

In terms of our base model, common initial belief in sequential rationality corresponds to the condition that a type respects common initial belief in the event that types initially believe in rationality at all information sets. The epistemic independence condition cannot be translated that easily into our base model. The problem is that the base model only allows for beliefs conditional on *specific* events, namely events in which some information set is reached. On the other hand, in order to formalize the epistemic independence condition we need to condition beliefs on more general events. There is, however, an important consequence of the epistemic independence condition that can be translated into our base model, namely that the event of reaching information set $h_i$ should not change player $i$'s belief about the actions and beliefs of players that did not precede $h_i$. In order to see this, choose a player $j$ that precedes $h_i$ and a player $k$ that does not precede $h_i$. Note that the event of player $j$ choosing the action leading to

$h_i$ is an event about player $j$, and that the event of player $k$ choosing a certain action and having a certain belief vector is an event about player $k$. Hence, epistemic independence says that player $i$'s belief about player $k$'s action and beliefs should not depend on whether player $j$ has moved the game towards $h_i$ or not. Moreover, it is exactly this consequence of epistemic independence that drives Stalnaker's backward induction result. In particular, if player $i$ initially believes that player $k$ chooses rationally at his information set, then player $i$ should continue to believe so if he observes that $h_i$ has been reached. If we drop the assumption that every player only controls one information set, the condition amounts to saying that a player should never revise his belief about the actions and beliefs at future and parallel information sets. Together with the condition of common initial belief in sequential rationality, this implies common belief in the event that types initially believe in rationality at all information sets. (See the discussion of Aumann's model for a similar argument). In terms of our base model, Stalnaker's sufficient condition for backward induction can thus be stated as follows:

**Stalnaker's condition BR:** Type $t_i$ respects common belief in the events that (1) types hold preferences as specified by $(\tilde{P}_i)_{i \in I}$, (2) types initially believe in rationality at all information sets, and (3) types do not change their belief about the opponents' choices and beliefs at future and parallel information sets.

### 3.11. Summary

The discussion of the various models and sufficient conditions for backward induction can be summarized by Table 1. The table shows that several sufficient conditions for backward induction, although formulated in completely different epistemic models, become equivalent once they have been translated into the language of our base model. Note also that there is no model assuming common belief in the events that (1) types always believe in rationality at *all* information sets, and (2) types never revise their beliefs about the opponents' preferences over terminal nodes. This is no surprise, since the papers by Reny (1992, 1993) have illustrated that these two events are in general incompatible. Perea's model maintains condition (1) and weakens condition (2), while the other models maintain condition (2) and weaken condition (1). Finally observe that all models assume (at least) common belief in the event that types initially believe in rationality at all information sets, plus some extra conditions on the players' belief revision procedures. If one would only assume the former, this would lead to the concept of *common certainty of rationality at the beginning of the game,* as defined by Ben-Porath (1997). This concept is considerably weaker than backward induction, as it may not even lead to the backward induction outcome. Hence, additional conditions on the players' belief revision policies are needed in each model to arrive at backward induction.

| | Ash | A&P | Aum | B&W | Cla | Fei1 | Fei2 | Per | Que | Sam | Sta |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Common belief in event that types ...** | | | | | | | | | | | |
| ... initially believe in rat. at all inf. sets | | | x | | | | | | | | x |
| ... always believe in rat. at future inf. sets that are believed to be reached | | | | | | | | | x | | |
| ... always believe in rat. at all future inf. sets | x | | | | | | | | | | |
| ... always believe in rat. at all future and parallel inf. sets | | x | | | | x | | | | | |
| ... always believe in rat. at all inf. sets | | | | | | | | x | | | |
| **Forward belief in substantive rat.** | | | | x | x | | x | | | | |
| **Forward belief in material rat.** | | | | | | | | | | x | |
| **Common belief in event that types ...** | | | | | | | | | | | |
| ... never revise belief about opponents' pref. relations | x | x | x | x | x | x | x | | x | x | x |
| ... do not revise belief about opponents' choices and beliefs at fut. and par. inf. sets | | | x | x | | | | | | | x |
| ... minimally revise belief about opponents' preferences and beliefs | | | | | | | | x | | | |
| ... hold point-beliefs | | | | | | | | x | x | | |

Table 1: Overview of sufficient conditions for backward induction

## References


[1] Asheim, G.B. (2002), On the epistemic foundation for backward induction, *Mathematical Social Sciences* **44,** 121-144.

[2] Asheim, G.B. and A. Perea (2005), Sequential and quasi-perfect rationalizability in extensive games, *Games and Economic Behavior* **53,** 15-42.

[3] Aumann, R. (1995), Backward induction and common knowledge of rationality, *Games and Economic Behavior* **8,** 6-19.

[4] Aumann, R. (1998), On the centipede game, *Games and Economic Behavior* **23,** 97-105.

[5] Balkenborg, D. and E. Winter (1997), A necessary and sufficient epistemic condition for playing backward induction, *Journal of Mathematical Economics* **27,** 325-345.

[6] Battigalli, P. (1997), On rationalizability in extensive games, *Journal of Economic Theory* **74,** 40-61.

[7] Battigalli, P. and M. Siniscalchi (2002), Strong belief and forward induction reasoning, *Journal of Economic Theory* **106,** 356-391.

[8] Ben-Porath, E. (1997), Rationality, Nash equilibrium and backwards induction in perfect-information games, *Review of Economic Studies* **64,** 23-46.

[9] Binmore, K. (1987), Modeling rational players, Part I, *Economics and Philosophy* **3,** 179-214.

[10] Bonanno, G. (1992), Rational beliefs in extensive games, *Theory and Decision* **33,** 153-176.

[11] Brandenburger, A., Friedenberg, A. and H.J. Keisler (2004), Admissibility in games, Available at http://pages.stern.nyu.edu/~abranden/

[12] Broome, J. and W. Rabinowicz (1999), Backwards induction in the centipede game, *Analysis* **59,** 237-242.

[13] Carroll, J.W. (2000), The backward induction argument, *Theory and Decision* **48,** 61-84.

[14] Clausing, T. (2003), Doxastic conditions for backward induction, *Theory and Decision* **54,** 315-336.

[15] Clausing, T. (2004), Belief revision in games of perfect information, *Economics and Philosophy* **20,** 89-115.

[16] Feinberg, Y. (2005), Subjective reasoning - dynamic games, *Games and Economic Behavior* **52,** 54-93.

[17] Pearce, D.G. (1984), Rationalizable strategic behavior and the problem of perfection, *Econometrica* **52,** 1029-1050.

[18] Perea, A. (2005), Minimal belief revision leads to backward induction, Available at http://www.personeel.unimaas.nl/a.perea/

[19] Priest, G. (2000), The logic of backward inductions, *Economics and Philosophy* **16,** 267-285.

[20] Quesada, A. (2002), Belief system foundations of backward induction, *Theory and Decision* **53,** 393-403.

[21] Quesada, A. (2003), From common knowledge of rationality to backward induction, *International Game Theory Review.*

[22] Reny, P.J. (1992), Rationality in extensive-form games, *Journal of Economic Perspectives* **6,** 103-118.

[23] Reny, P.J. (1993), Common belief and the theory of games with perfect information, *Journal of Economic Theory* **59,** 257-274.

[24] Rosenthal, R.W. (1981), Games of perfect information: predatory pricing and the chain-store paradox, *Journal of Economic Theory* **25,** 92-100.

[25] Rubinstein, A. (1991), Comments on the interpretation of game theory, *Econometrica* **59,** 909-924.

[26] Samet, D. (1996), Hypothetical knowledge and games with perfect information, *Games and Economic Behavior* **17,** 230-251.

[27] Stalnaker, R. (1996), Knowledge, belief and counterfactual reasoning in games, *Economics and Philosophy* **12,** 133-163.

[28] Stalnaker, R. (1998), Belief revision in games: forward and backward induction, *Mathematical Social Sciences* **36,** 31-56.

[29] Zermelo, E. (1913), Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels, *Proceedings Fifth International Congress of Mathematicians* **2,** 501-504.